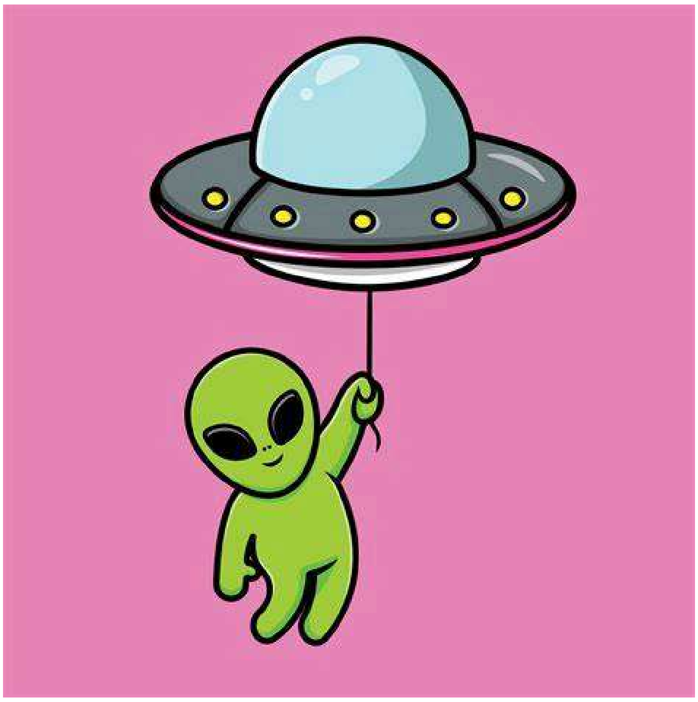
Project assignment. Statistical programming fundamentals.

Jaime Rodriguez Roncero

2025-01-31

In this analysis, we explored a dataset of reported UFO sightings, applying data cleaning, transformation, and visualization techniques to uncover key patterns. We examined the distribution of sightings across countries, analyzed trends over time and identified the most frequently reported UFO shapes.



1. Data Preparation

1.1. Data import

Loading our dataset.

```
rm(list=ls())
df <- read_csv("../data/01_initial_dataset.csv")
```

1.2. Initial exploration.

View data structure.

```
head(df)
```

```
## # A tibble: 6 × 11
##   datetime city  state country shape `duration (seconds)` `duration (hours/min)`
##   <chr>    <chr> <chr> <chr>   <chr>                <dbl> <chr>
## 1 10/10/1… san … tx    us      cyli…                 2700 45 minutes
## 2 10/10/1… lack… tx    <NA>    light                 7200 1-2 hrs
## 3 10/10/1… ches… <NA>  gb      circ…                   20 20 seconds
## 4 10/10/1… edna  tx    us      circ…                   20 1/2 hour
## 5 10/10/1… kane… hi    us      light                  900 15 minutes
## 6 10/10/1… bris… tn    us      sphe…                  300 5 minutes
## # ℹ 4 more variables: comments <chr>, `date posted` <chr>, latitude <chr>,
## #   longitude <chr>
```

```
glimpse(df)
```

```
## Rows: 88,875
## Columns: 11
## $ datetime               <chr> "10/10/1949 20:30", "10/10/1949 21:00", "10/10/…
## $ city                   <chr> "san marcos", "lackland afb", "chester (uk/engl…
## $ state                  <chr> "tx", "tx", NA, "tx", "hi", "tn", NA, "ct", "al…
## $ country                <chr> "us", NA, "gb", "us", "us", "us", "gb", "us", "…
## $ shape                  <chr> "cylinder", "light", "circle", "circle", "light…
## $ `duration (seconds)`   <dbl> 2700, 7200, 20, 20, 900, 300, 180, 1200, 180, 1…
## $ `duration (hours/min)` <chr> "45 minutes", "1-2 hrs", "20 seconds", "1/2 hou…
## $ comments               <chr> "This event took place in early fall around 194…
## $ `date posted`          <chr> "4/27/2004", "12/16/2005", "1/21/2008", "1/17/2…
## $ latitude               <chr> "29.8830556", "29.38421", "53.2", "28.9783333",…
## $ longitude              <chr> "-97.9411111", "-98.581082", "-2.916667", "-96.…
```

We see that each row correspond to a different UFO sighting and contains 11 columns with information about it (country, time,…)

1.3 Data cleaning.

Lets start by removing the following columns:

- *duration (hours/min).* It has irrelevant information having *duration (seconds)* column and it has inconsistent data format (1/2 hour, 1-2 hours, 20 seconds).
- *date posted.* Does not contain relevant data for the analysis we will perform.
- *city.* It is very specific (for the same city each row has different annotations). We will use latitude and longitude if we need this kind of data.

```
df$`duration (hours/min)` <- NULL
df$`date posted` <- NULL
df$city <- NULL
```

Transform column's data types to something more suitable for them:

```
df$datetime <- as.POSIXct(df$datetime, format = "%m/%d/%Y %H:%M")
df$state <- as.factor(df$state)
df$country <- as.factor(df$country)
df$shape <- as.factor(df$shape)
df$longitude <- as.numeric(df$longitude)
df$latitude <- as.numeric(df$latitude)

df %>% rename(duration = `duration (seconds)`) -> df
```

Look at summary statistics to catch any anomalies (e.g., impossible durations, NA's, etc.).

```
summary(df)
```

```
##     datetime                          state       country          shape      
##  Min.   :1906-11-11 00:00:00.00   ca     :10450   au  :  593   light   :17872  
##  1st Qu.:2001-06-04 17:30:00.00   wa     : 4653   ca  : 3266   triangle: 8489  
##  Median :2006-09-21 20:00:00.00   fl     : 4598   de  :  112   circle  : 8453  
##  Mean   :2004-03-04 03:23:10.56   tx     : 4050   gb  : 2050   fireball: 6562  
##  3rd Qu.:2011-05-08 22:30:00.00   ny     : 3511   us  :70293   unknown : 6319  
##  Max.   :2014-05-08 18:45:00.00   (Other):54094   NA's:12561   (Other) :38062  
##  NA's   :2                        NA's   : 7519                NA's    : 3118  
##     duration          comments            latitude        longitude      
##  Min.   :       0   Length:88875       Min.   :-82.86   Min.   :-176.66  
##  1st Qu.:      15   Class :character   1st Qu.: 34.03   1st Qu.:-112.07  
##  Median :     120   Mode  :character   Median : 39.23   Median : -87.65  
##  Mean   :    8373                      Mean   : 37.45   Mean   : -85.02  
##  3rd Qu.:     600                      3rd Qu.: 42.72   3rd Qu.: -77.77  
##  Max.   :97836000                      Max.   : 72.70   Max.   : 178.44  
##  NA's   :5                             NA's   :197      NA's   :196      
```

We note the following things:

- Max of *datatime* column is 2014-05-08. This means data acquisition stop before finishing last year.
- Values for latitude and longitude are withing the allowed range ([-90,90] and [-180,180]).
- State and country columns have a large number of NA's. We will allow this because there are levels for only 5 countries and state field for no U.S. countries must be blank.
- We have outliers in duration column.

First lets remove all the rows containing NA's that are not in state and country columns (We have 88 thousand rows and the rows removed are a few hundreds, so we can take this.)

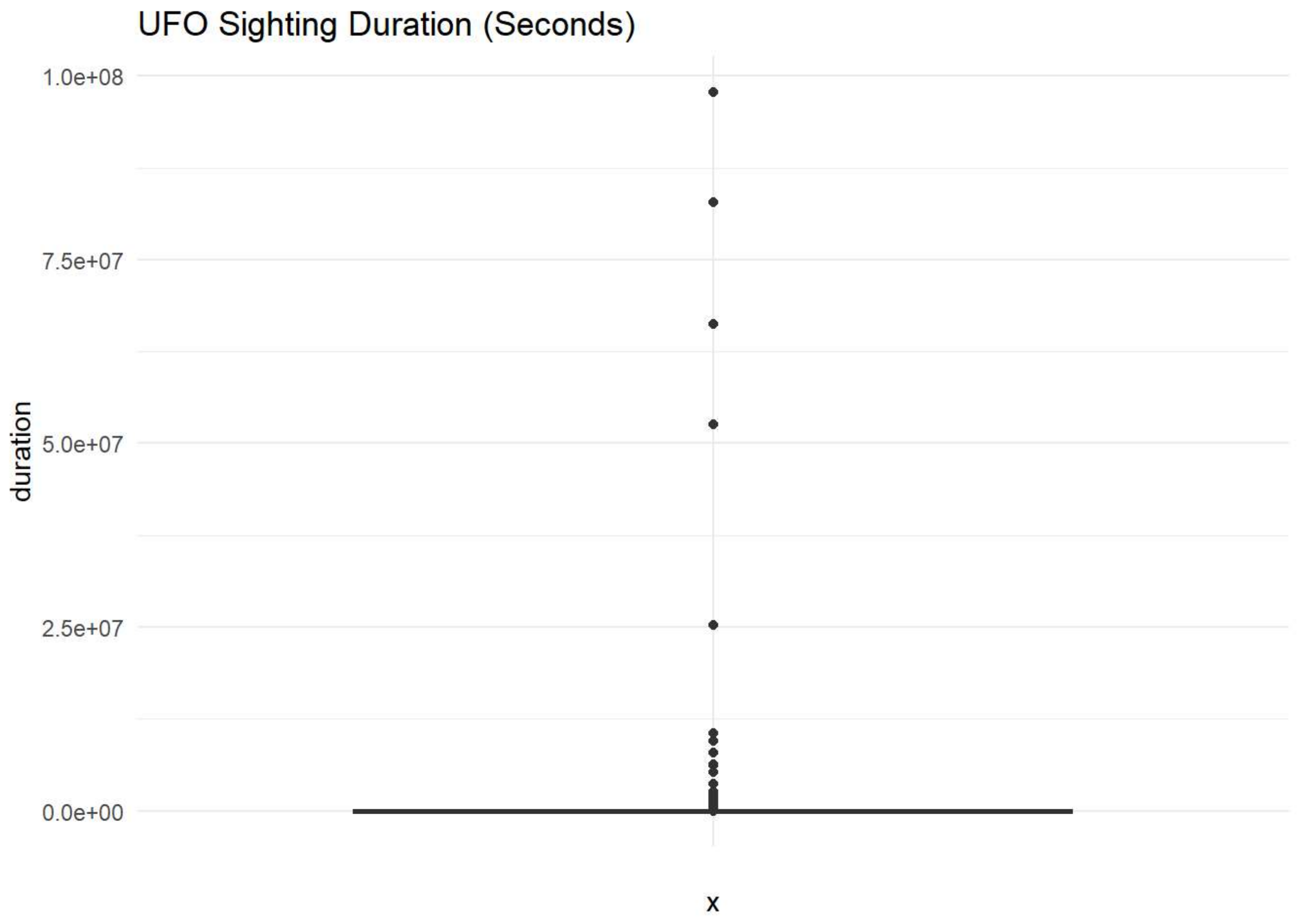
```
df <- df %>% filter( if_all( .cols = -c(state, country), ~ !is.na(.) ) )
```

```
summary(df)
```

```
##     datetime                          state            country           shape      
##  Min.   :1906-11-11 00:00:00.00   ca     :10086   au  :  577   light   :17870  
##  1st Qu.:2001-08-24 22:15:00.00   fl     : 4460   ca  : 3207   triangle: 8488  
##  Median :2006-11-18 22:30:00.00   wa     : 4242   de  :  110   circle  : 8451  
##  Mean   :2004-04-29 14:44:49.33   tx     : 3959   gb  : 1992   fireball: 6562  
##  3rd Qu.:2011-06-15 15:10:00.00   ny     : 3413   us  :68053   unknown : 6318  
##  Max.   :2014-05-08 18:45:00.00   (Other):52435   NA's:11804   other   : 6247  
##                                   NA's   : 7148                (Other) :31807  
##     duration          comments           latitude        longitude      
##  Min.   :       0   Length:85743       Min.   :-82.86   Min.   :-176.66  
##  1st Qu.:      15   Class :character   1st Qu.: 34.02   1st Qu.:-111.92  
##  Median :     120   Mode  :character   Median : 39.19   Median : -87.65  
##  Mean   :    7542                      Mean   : 37.43   Mean   : -84.91  
##  3rd Qu.:     600                      3rd Qu.: 42.66   3rd Qu.: -77.72  
##  Max.   :97836000                      Max.   : 72.70   Max.   : 178.44  
##
```

Now lets handle outliers in duration:

```
ggplot(df, aes(x = "", y = duration)) +
  geom_boxplot() +
  labs(
    title = "UFO Sighting Duration (Seconds)"
  ) +
  theme_minimal()
```

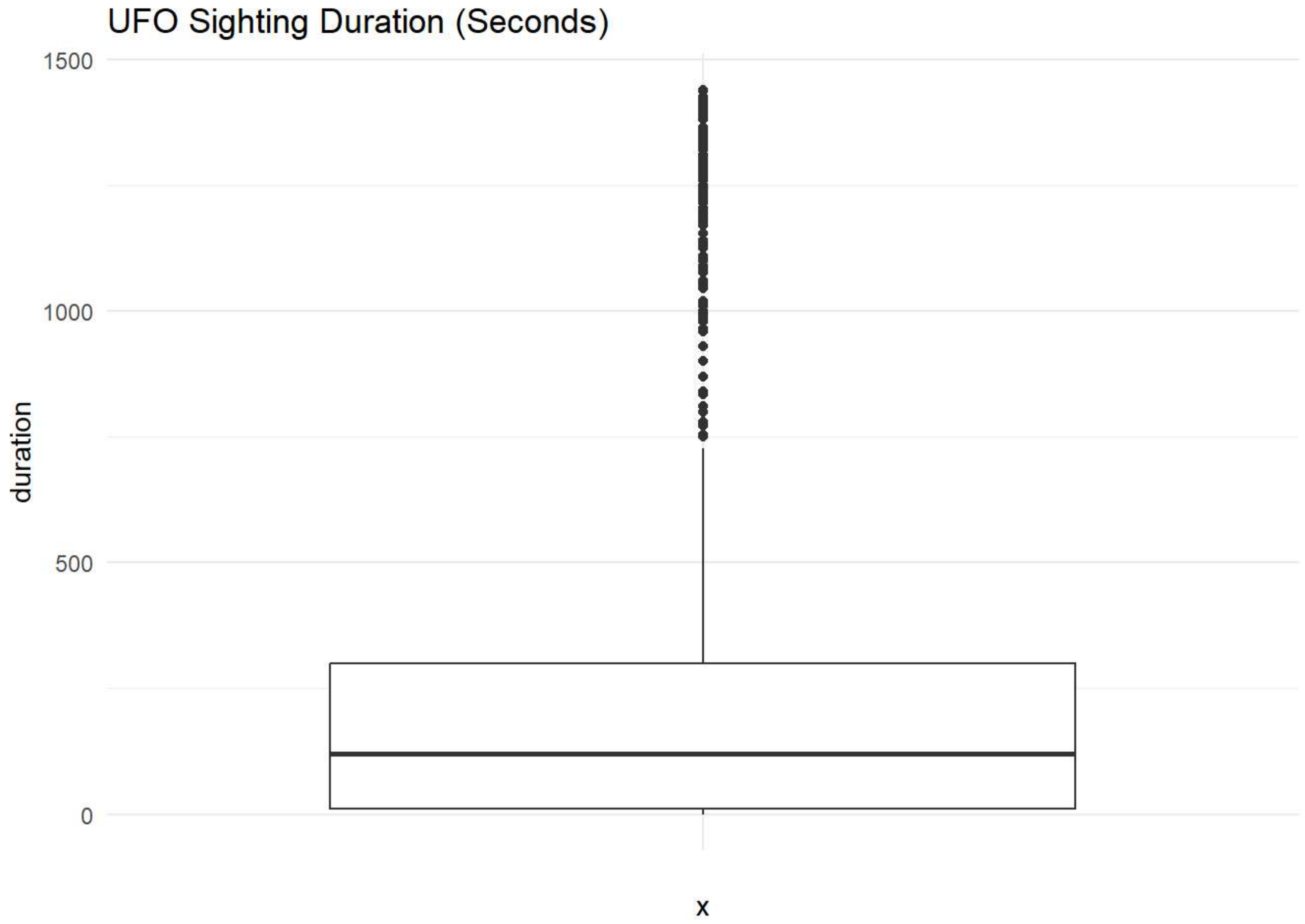


If we take a look to this outliers, they do not have any type of special information or common pattern, so we will remove them using the 1.5IQR rule.

```
df %>% filter(duration > 2.5e7) %>% print
```

```
## # A tibble: 6 × 8
##   datetime            state country shape  duration comments  latitude longitude
##   <dttm>              <fct> <fct>   <fct>     <dbl> <chr>        <dbl>     <dbl>
## 1 1983-10-01 17:00:00 <NA>  gb      sphere 97836000 Firstly&…     52.5     -1.92
## 2 1969-06-30 22:45:00 <NA>  gb      cone   25248000 First ti…     51.1     -3
## 3 2010-06-03 23:30:00 on    ca      other  82800000 ((HOAX??…     45.4    -75.7
## 4 2012-08-10 21:00:00 wa    us      light  52623200 There ha…     46.2   -119.
## 5 2002-08-24 01:00:00 fl    us      light  52623200 bright s…     27.0    -82.4
## 6 1991-09-15 18:00:00 ar    us      light  66276000 Orange o…     35.2    -92.4
```

```
Q1 <- quantile(df$duration, 0.25)
Q3 <- quantile(df$duration, 0.75)
IQR_value <- Q3 - Q1

lower_bound <- Q1 - 1.5 * IQR_value
upper_bound <- Q3 + 1.5 * IQR_value

df <- df %>%
  filter(duration >= lower_bound, duration <= upper_bound)

ggplot(df, aes(x = "", y = duration)) +
  geom_boxplot() +
  labs(
    title = "UFO Sighting Duration (Seconds)"
  ) +
  theme_minimal()
```



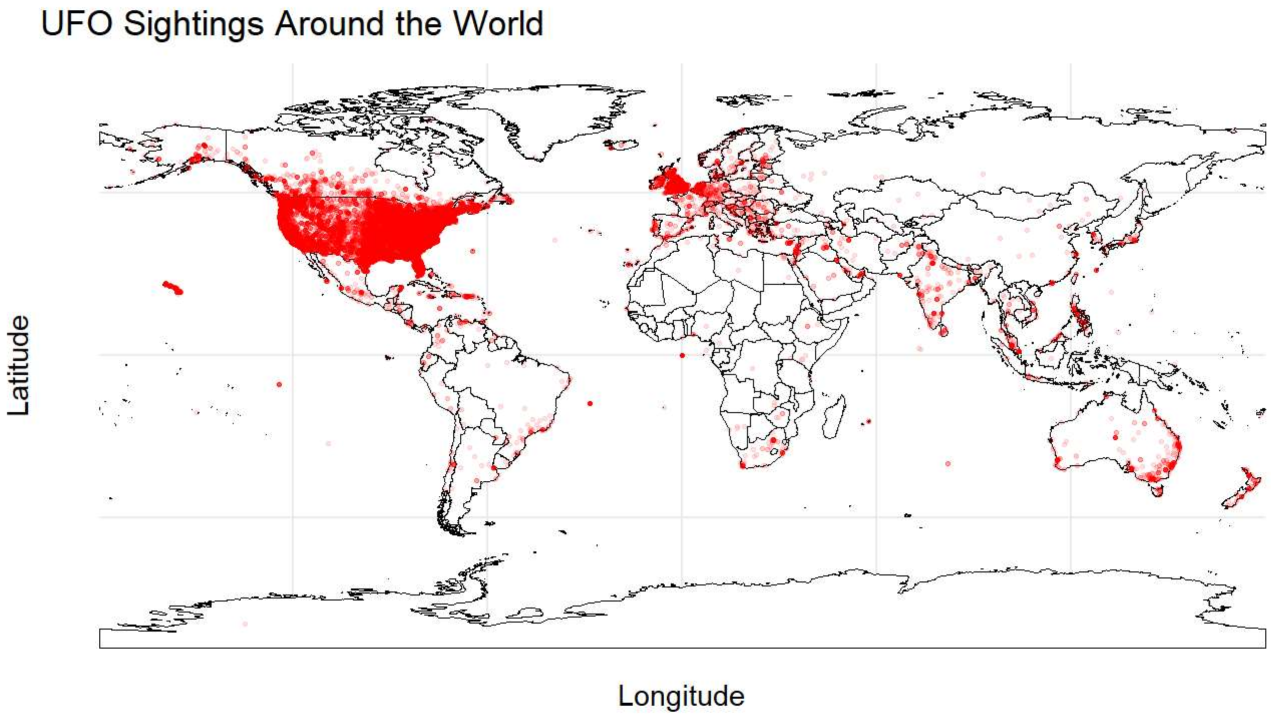
We are now ready for analysis.

```
write_csv(df, "../data/02_cleaned_dataset.csv")
```

2. Geographical analysis: Which locations report the highest number of UFO sightings?

Lets begin with a general view of sighting spots around the world:

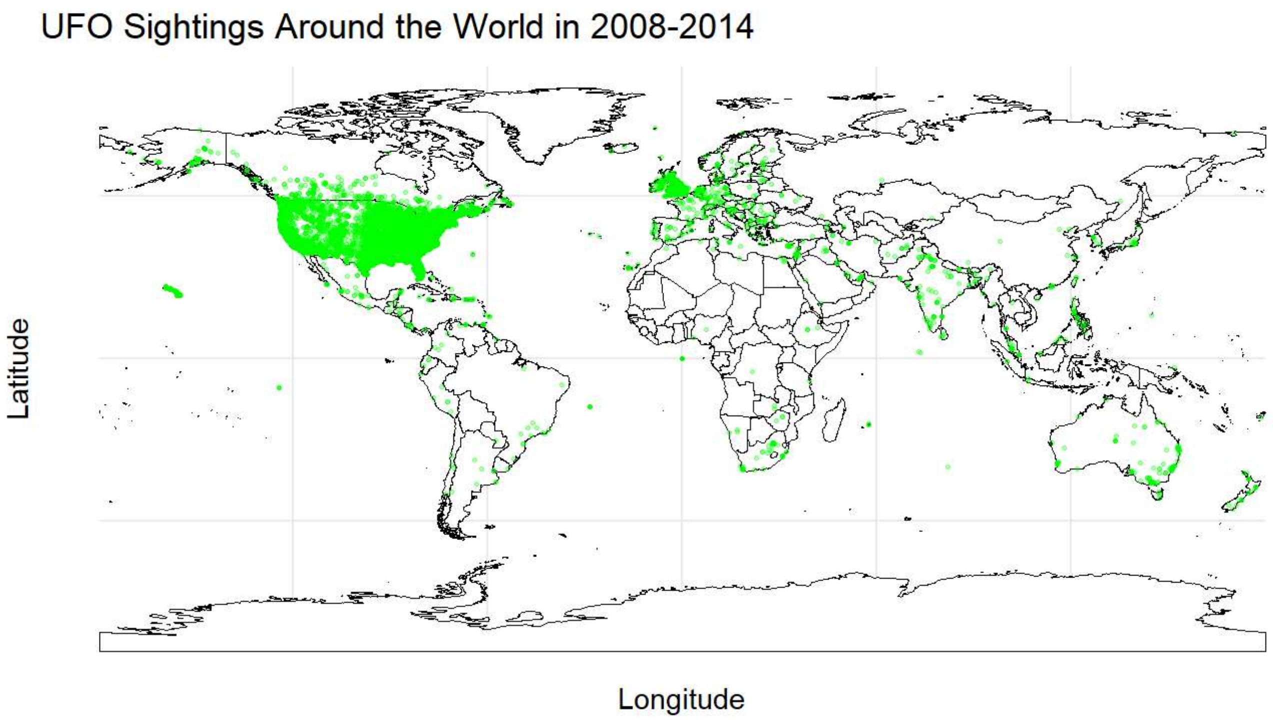
```
world <- ne_countries(scale = "medium", returnclass = "sf")

ggplot() +
  geom_sf(data = world, fill = "white", color = "black", size = 0.2) +
  geom_point(
    data = df,
    aes(x = longitude, y = latitude),
    color = "red",
    alpha = 0.1,
    size = 0.5
  ) +
  coord_sf() +
  labs(
    title = "UFO Sightings Around the World",
    x = "Longitude",
    y = "Latitude"
  ) +
  theme_minimal()
```



We observe that a big part of the data comes from U.S, Europe and Australia while other crowded countries have few reports (Brazil, India, China,…) This suggest that access to global communications has a big importance for sightings to be reported (or that aliens are not interested in visiting Africa.)

Lets do the same analysis, but only having into account reports from 2008, when Internet is accessible globally:

```
df$year <- year(df$datetime)

ggplot() +
  geom_sf(data = world, fill = "white", color = "black", size = 0.2) +
  geom_point(
    data = filter(df, year >= 2008),
    aes(x = longitude, y = latitude),
    color = "green",
    alpha = 0.3,
    size = 0.5
  ) +
  coord_sf() +
  labs(
    title = "UFO Sightings Around the World in 2008-2014",
    x = "Longitude",
    y = "Latitude"
  ) +
  theme_minimal()
```



Results are very similar. It is curious that crowded countries with an easy access to internet like Argentina or Mexico do not have so many reports. This lead us to think that there is a cultural reason in U.S. that makes people more prone to identify flying objects as UFOs or that, if aliens really exist, they love to visit this country.

Now we are comparing reports per inhabitant for the countries we have data:

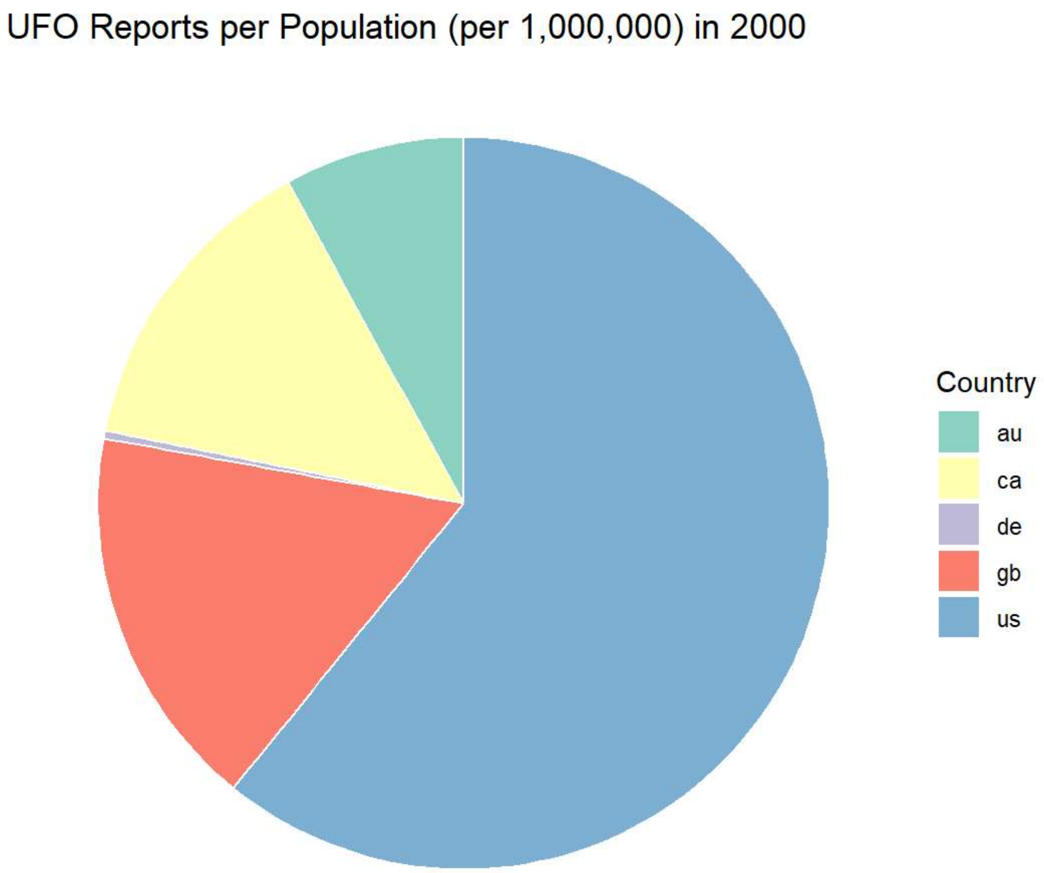
```
data(pop)
countries_of_interest <- c("Australia", "Canada", "Germany", "United Kingdom", "United States of America")
pop_extract <- pop %>% filter(name %in% countries_of_interest) %>% select(name, '2000')
print(pop_extract)
```

```
##                       name      2000
## 1                Australia  18991.43
## 2           United Kingdom  58923.31
## 3                  Germany  81400.88
## 4                   Canada  30588.38
## 5 United States of America 281710.91
```

```
df_pop <- data.frame(table(filter(df, !is.na(country))$country))
df_pop %>% mutate(freq_norm = Freq / pop_extract$'2000' * 1e6)  -> df_pop # Normalizing using 2000's populations
df_pop$freq_norm %>% round -> df_pop$freq_norm
print(df_pop)
```

```
##   Var1  Freq freq_norm
## 1   au   529     27855
## 2   ca  2853     48419
## 3   de    97      1192
## 4   gb  1819     59467
## 5   us 59902    212636
```

```
ggplot(df_pop, aes(x = "", y = freq_norm, fill = Var1)) +
  geom_bar(stat = "identity", width = 1, color = "white") +
  coord_polar("y") +
  scale_fill_brewer(palette = "Set3") +
  theme_void() +
  labs(
    fill = "Country",
    title = "UFO Reports per Population (per 1,000,000) in 2000"
  )
```



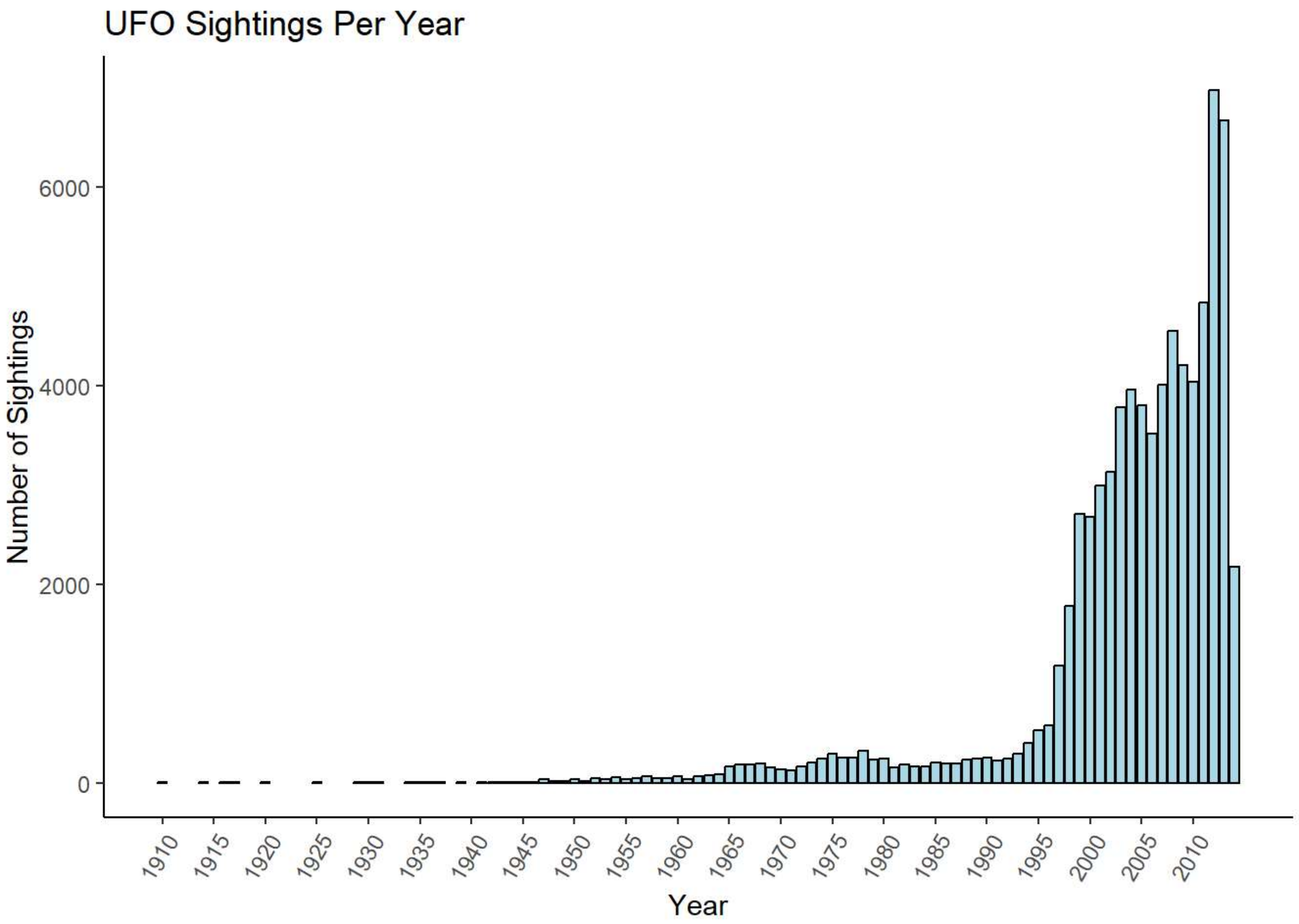
UFO Reports per Population (per 1,000,000) in 2000

The plot shows that reports per inhabitant is significantly bigger in U.S than other countries. This confirms our previous hypothesis.

3. Temporal analysis. Is there a specific time of year when these events are more common?

We start analyzing the total number of reports through the years:

```
ggplot(df, aes(x = year)) +
  geom_bar(fill = "lightblue", color = "black") +
  labs(title = "UFO Sightings Per Year", x = "Year", y = "Number of Sightings") +
  scale_x_continuous(
    breaks = seq(min(df$year, na.rm = TRUE), max(df$year, na.rm = TRUE), by = 5)
  ) +
  theme_classic() +
  theme(
    axis.text.x = element_text(angle = 60, hjust = 1)
  )
```

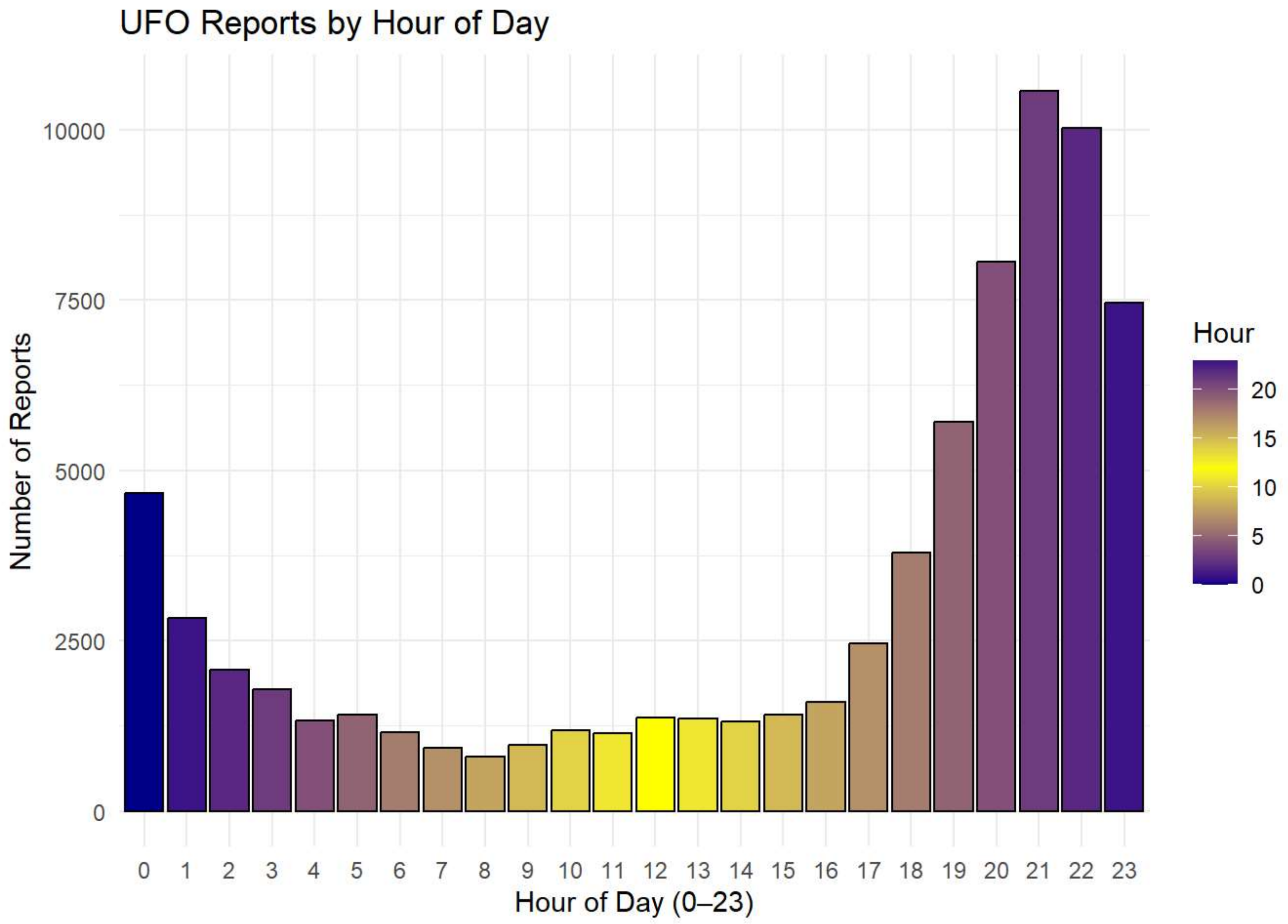


We can observe that sightings increases over years and the growth rate also rises. Here are my hypothesis for different periods:

- (1906 - 1946) The number is very small. Limited technology and communication channels resulted in fewer reported sightings.
- (1946 - 1994) Grows slowly. Increased media coverage (e.g., 1947 Roswell incident, sci-fi films) and the Cold War era heightened public interest and awareness.
- (1994 - 2014) Grows fast. Widespread internet access, mobile phones, and social media made reporting sightings easier and more accessible, leading to rapid growth. Last year is lower because the data acquisition stopped during the year.

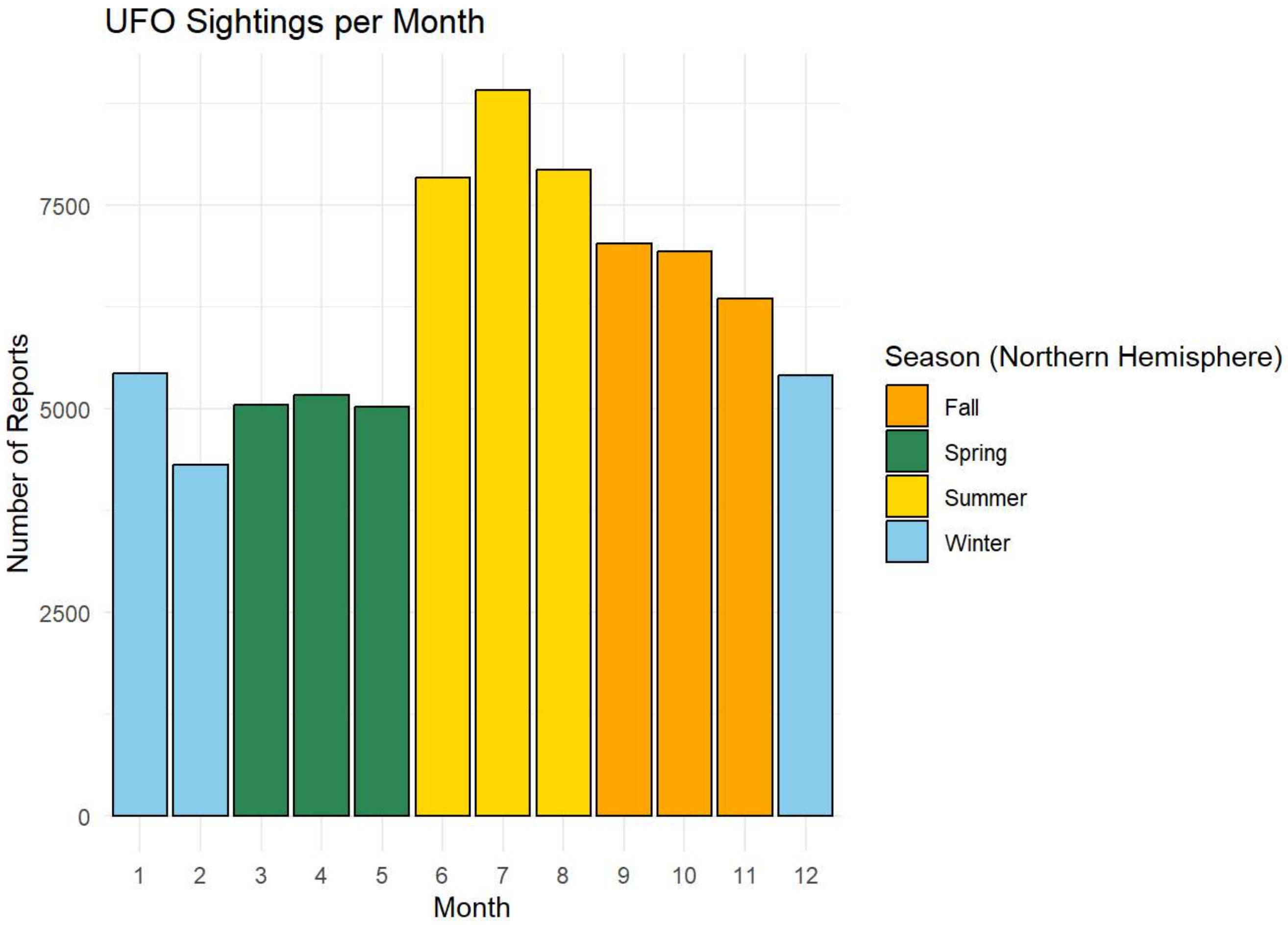
Looking at sightings hours could give us some important information:

```
df_by_hour <- df %>%
  mutate(hour = hour(datetime)) %>%       # integer from 0 to 23
  group_by(hour) %>%
  summarize(n_reports = n(), .groups = "drop")

ggplot(df_by_hour, aes(x = factor(hour), y = n_reports, fill = hour)) +
  geom_bar(stat = "identity", color = "black") +
  scale_fill_gradient2(
    low = "darkblue",
    mid = "yellow",
    high = "darkblue",
    midpoint = 12,
    name = "Hour"
  ) +
  labs(
    x = "Hour of Day (0–23)",
    y = "Number of Reports",
    title = "UFO Reports by Hour of Day"
  ) +
  theme_minimal()
```



It is curious to observe that there are more reports in the hours when people usually have free time (After job and before bed).

```r
df_by_month <- df %>%
  mutate(
    month = month(datetime),  # integer 1..12
    season = case_when(
      month %in% c(12, 1, 2)  ~ "Winter",
      month %in% c(3, 4, 5)   ~ "Spring",
      month %in% c(6, 7, 8)   ~ "Summer",
      month %in% c(9, 10, 11) ~ "Fall"
    )
  ) %>%
  group_by(month, season) %>%
  summarize(n_sightings = n(), .groups = "drop")

ggplot(df_by_month, aes(x = factor(month), y = n_sightings, fill = season)) +
  geom_bar(stat = "identity", color = "black") +
  scale_fill_manual(
    values = c(
      "Winter" = "skyblue",
      "Spring" = "seagreen",
      "Summer" = "gold",
      "Fall"   = "orange"
    )
  ) +
  labs(
    x = "Month",
    y = "Number of Reports",
    fill = "Season (Northern Hemisphere)",
    title = "UFO Sightings per Month"
  ) +
  theme_minimal()
```

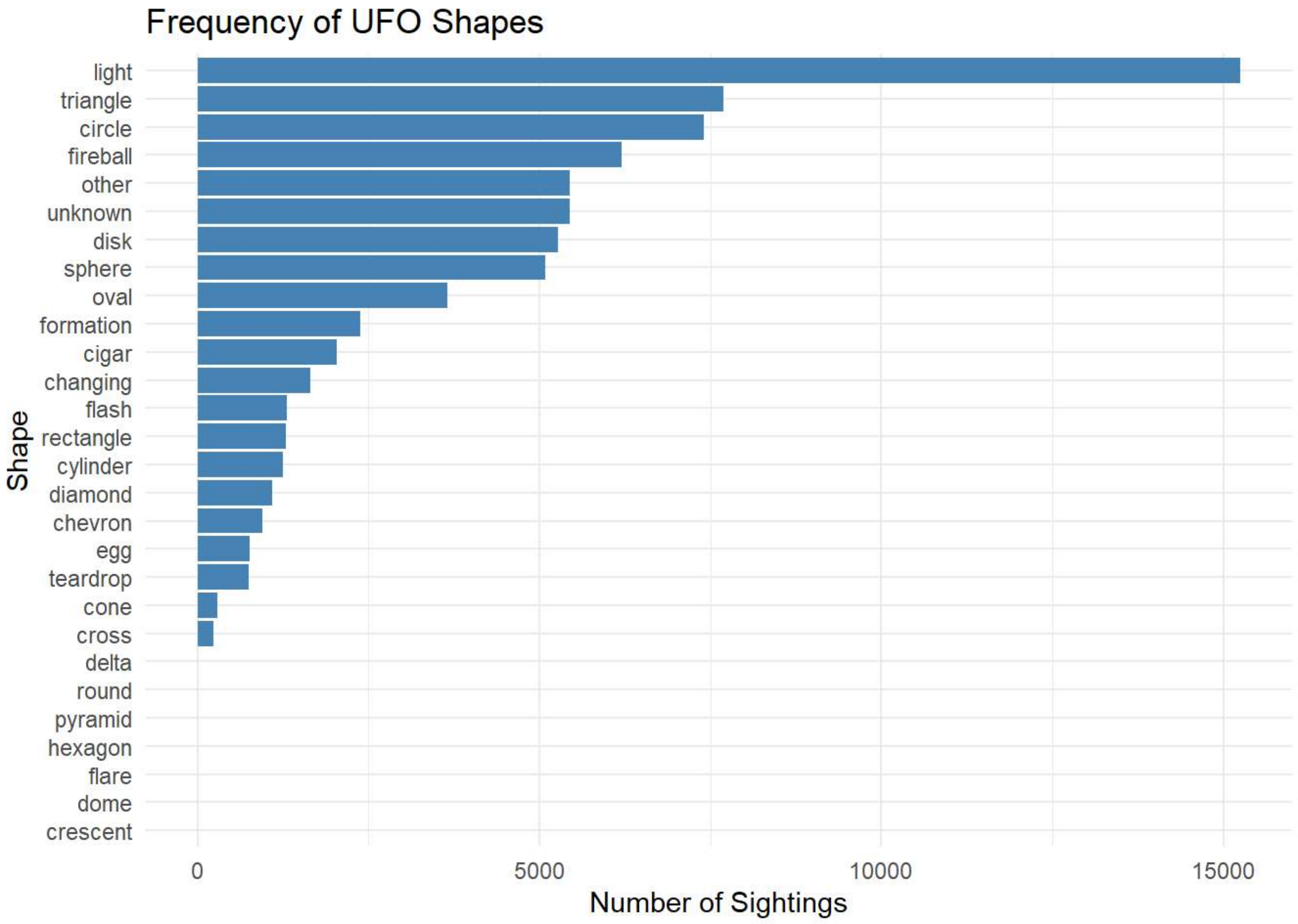


The seasons with more reports is summer (Taking into account that data comes mainly from Northern Hemisphere: U.S. and Europe). The hypothesis here is that people are often outdoors, so chances of an event to be reported are higher. Now we know that if we want to look for UFOs, the best time for doing it is on summer over 9pm.

4. Are there recurring patterns in the shapes?

In this case would be interesting to do frequency analysis over *comments* column using packages like *udpipe* to find the most common words appearing in reports descriptions, but unfortunately I do not have the hardware to perform this task. We would analize instead *shapes* column:

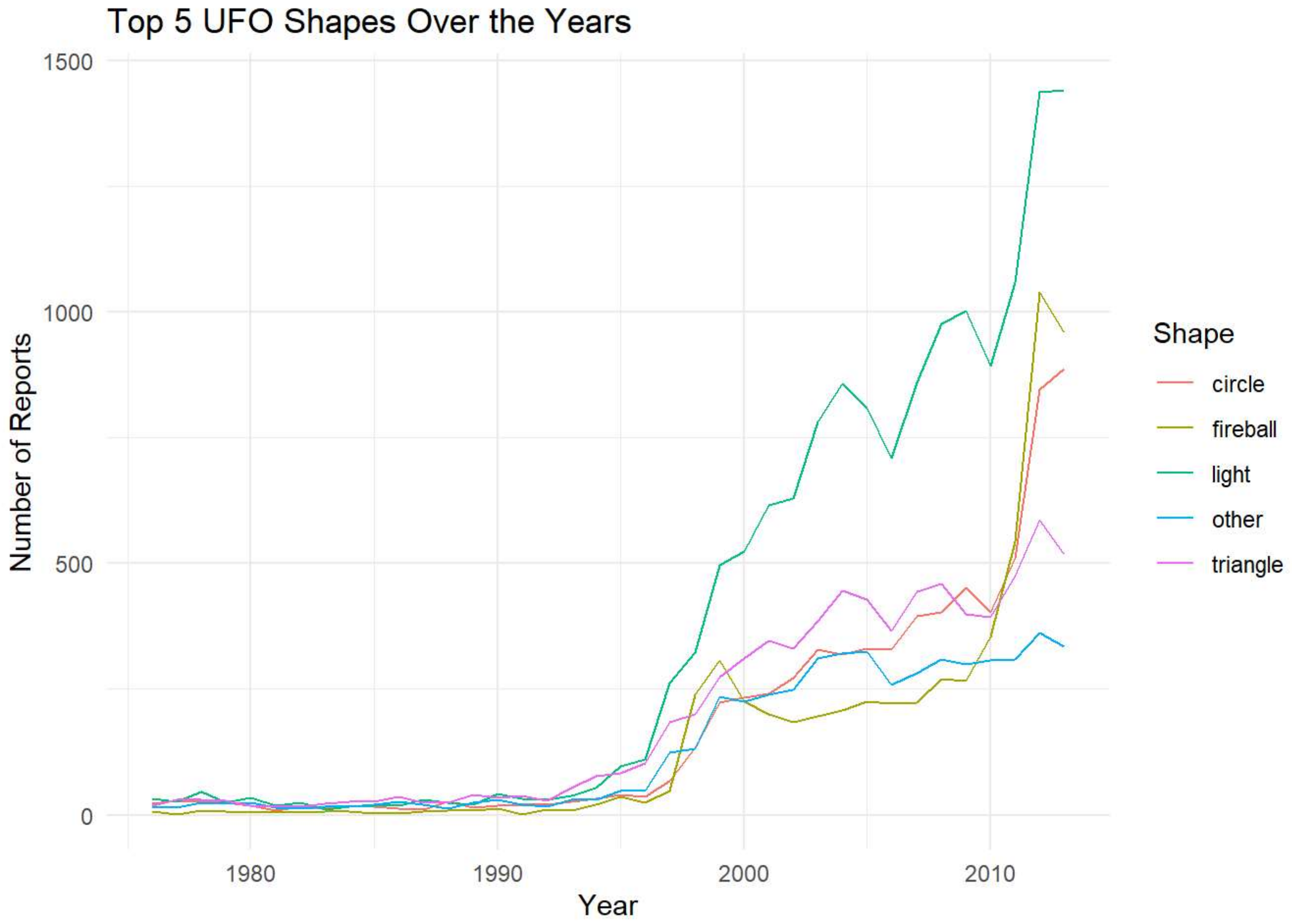
```
shape_freq <- df %>%
  filter(!is.na(shape)) %>%    # Remove rows with missing shape
  group_by(shape) %>%
  summarize(n = n(), .groups = "drop") %>%
  arrange(desc(n))             # Sort descending by frequency

ggplot(shape_freq, aes(x = reorder(shape, n), y = n)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(
    x = "Shape",
    y = "Number of Sightings",
    title = "Frequency of UFO Shapes"
  ) +
  theme_minimal()
```



```
top5_shapes <- shape_freq %>%
  slice_head(n = 5) %>%
  pull(shape)

df_yearly_shapes <- df %>%
  filter(
    shape %in% top5_shapes,
    year > 1975,
    year < 2014
  ) %>%
  group_by(shape, year) %>%
  summarize(n = n(), .groups = "drop")

head(df_yearly_shapes)
```

```
## # A tibble: 6 × 3
##   shape   year     n
##   <fct>  <dbl> <int>
## 1 circle  1976    22
## 2 circle  1977    29
## 3 circle  1978    26
## 4 circle  1979    25
## 5 circle  1980    18
## 6 circle  1981     9
```

```
ggplot(df_yearly_shapes, aes(x = year, y = n, color = shape)) +
  geom_line() +
  labs(
    title = "Top 5 UFO Shapes Over the Years",
    x = "Year",
    y = "Number of Reports",
    color = "Shape"
  ) +
  theme_minimal()
```



Description of the events as *light* has increased over the years, being with difference the most common description in 20th century. In my opinion, the development of aviation has a lot to say about this.

5. Conclusions.

Our exploratory analysis of the UFO sightings dataset reveals several patterns. First, the frequency of reports has increased over time, although part of this trend likely reflects greater public awareness and reporting mechanisms in more recent years. Geographically, the United States accounts for the largest volume of documented sightings, with other English-speaking countries (such as Canada, the United Kingdom, and Australia) also showing significant activity. When examining shapes, “light,” “triangle,” and “circle” are among the most commonly reported forms, suggesting certain recurring visual perceptions. Additionally, the analysis of sightings throughout the day reveals a peak in reports during the late evening hours, possibly correlating with times people are more likely to be outdoors or actively sky-watching. While these findings provide interesting insights into human observations and reporting behavior, they do not offer direct evidence of extraterrestrial origins; rather, they illustrate social and cultural factors influencing witness accounts, as well as the limitations of anecdotal data.